W160–W165 *Nucleic Acids Research, 2004, Vol. 32, Web Server issue*
DOI: 10.1093/nar/gkh418

The ERPIN server: an interface to profile-based RNA motif identification

André Lambert, Jean-Fred Fontaine[1], Matthieu Legendre[1], Fabrice Leclerc[2], Emmanuelle Permal[3], François Major[3], Harald Putzer[4], Olivier Delfour[5], Bernard Michot[5] and Daniel Gautheret[1,*]

CNRS UMR 6207 and [1]INSERM ERM 206, Université de la Méditerranée, Luminy Case 906, 13288 Marseille, Cedex 09, France, [2]UMR 7567 CNRS-UHP, Université Henri Poincaré, 54506 Vandoeuvre-lès-Nancy cedex, France, [3]Département d'Informatique et de Recherche Opérationnelle, Université de Montréal, CP 6128, Succ. Centre-Ville, Montréal, Québec, Canada H3C 3J7, [4]CNRS UPR 9073, 13 rue P. et M. Curie, 75005 Paris, France and [5]ACTiGenics, 10 avenue de l'Europe, 31525 Ramonville St Agne, France

Received February 13, 2004; Revised and Accepted April 5, 2004

ABSTRACT

ERPIN is an RNA motif identification program that takes an RNA sequence alignment as an input and identifies related sequences using a profile-based dynamic programming algorithm. ERPIN differs from other RNA motif search programs in its ability to capture subtle biases in the training set and produce highly specific and sensitive searches, while keeping CPU requirements at a practical level. In its latest version, ERPIN also computes *E*-values, which tell biologists how likely they are to encounter a specific sequence match by chance—a useful indication of biological significance. We present here the ERPIN online search interface (http://tagc.univ-mrs.fr/erpin/). This web server automatically performs ERPIN searches for different RNA genes or motifs, using predefined training sets and search parameters. With a couple of clicks, users can analyze an entire bacterial genome or a genomic segment of up to 5Mb for the presence of tRNAs, 5S rRNAs, SRP RNA, C/D box snoRNAs, hammerhead motifs, miRNAs and other motifs. Search results are displayed with sequence, score, position, *E*-value and secondary structure graphics. An example of a complete genome scan is provided, as well as an evaluation of run times and specificity/sensitivity information for all available motifs.

INTRODUCTION

The last few years have seen a continuous stream of novel noncoding RNA (ncRNA) genes and motifs being reported. Known RNA functions now display a wonderful diversity, ranging from genetic data storage to sensing, transport, targeting and even regulation of essential events such as cell differentiation or cell death. These discoveries have led biologists to undertake a systematic scrutiny of genomic sequences for functional ncRNAs in the form of either independent genes or structural motifs in the untranslated part of transcripts. This effort is carried out in part using experimental RNA amplification strategies. However, with the growth of sequence databases and the development of specific algorithms for RNA structure detection, bioinformatics is now emerging as an inexpensive yet efficient alternative. What tools are now available for computational RNA motif identification? Standard sequence alignment programs are generally not suited to ncRNA searches, because ncRNA is largely characterized by long-range base pair interactions and not by its linear sequence. Bioinformatics has addressed this problem in several ways, notably through descriptor-based systems in which the topology of base-paired regions is specified by the user (1–3), stochastic context free grammars (SCFGs), which use a complete statistical model of RNA elements (4,5), and secondary structure profiles, which are position weight matrices describing stems and single strands in the RNA motif, as implemented in the ERPIN program (6).

ERPIN takes an RNA sequence alignment and secondary structure annotation as input. From this 'training set', the

*To whom correspondence should be addressed. Tel: +33 491 828639; Fax: +33 491 828621; Email: gautheret@esil.univ-mrs.fr
Present address:
Jean-Fred Fontaine, INSERM EMI U 00.18, CHU d'Angers, 49033 Angers, France

program constructs a lod-score profile for each helix or single strand in the alignment, and dynamic programming is applied to identify significant occurrences of these profiles in any database sequence. ERPIN differs from descriptor-based RNA motif search programs in its ability to exploit the most subtle biases in helices or single strands, resulting in highly specific and sensitive searches. In addition, unlike other statistical approaches such as SCFG models (4,5), ERPIN handles pseudoknots and, more importantly, is able to perform very fast database scans and is thus more practical for genomic annotation purposes. These merits notwithstanding, developing a good RNA sequence alignment and finely tuning an ERPIN command line to achieve an accurate and efficient motif search remain an expert task. We thus felt that more biologists could benefit from the ERPIN program through a web interface, where all training sets and search parameters are predefined to ensure optimized search conditions. We present here the ERPIN web server, which enables biologists to scan DNA segments of up to 5 Mb for the presence of known RNA genes and motifs. There are currently more than 30 genes/motifs to search for, and this collection is quickly growing through a collaborative effort. This website is also the home of the ERPIN project, whence sources, tutorials, documentation and training sets can be retrieved.

RECENT IMPROVEMENTS IN THE ERPIN SOFTWARE

The current version of the ERPIN program is 4.0. Major revisions since version 1.0 have brought the following improvements.

Arbitrarily complex regions. ERPIN 1.0 was limited to RNA motifs containing an arrangement of one or two double helices (nested, consecutive or pseudoknotted). Larger motifs had to be defined using a combination of these basic elements, which turned out to be quite complicated when dealing with even mid-sized RNAs. Since version 2.0, ERPIN has been able to deal with regions of arbitrary length containing helices and single strands arranged in any order.

Multi-level search. Since searches involving multiple elements can quickly become prohibitive in terms of CPU time and memory usage, we implemented a multi-level search option. This allows the program to search motif parts in turn only after previous parts have been identified using a defined cutoff score.

E-values. Since version 3.9, ERPIN has also computed an Expect value (*E*-value). This statistical measure tells biologists how likely they are to encounter a specific sequence match by chance during their database search. An *E*-value of *N* for a hit of score *S* indicates that one should expect *N* hits of score *S* or above purely by chance. Of course, this value depends on the database size. *E*-values were first introduced into bioinformatics with the BLAST program (7), and they have been adopted since by all biologists in need of a convenient means of assessing the significance of a database hit. In ERPIN 4.0, *E*-values are computed based on a prior analysis of score distribution in the lod-score profiles. These computed *E*-values show a very good correlation with the actual numbers of hits obtained through simulations on random databases (A. Lambert, M. Legendre, J.-F. Fontaine and A. Gautheret, manuscript in preparation).

THE WEB SERVER

In spite of these improvements, conducting a database search using the standalone ERPIN program remains a challenging task. Not only is an accurate RNA sequence alignment required (which is usually best done by RNA structure specialists), but search parameters have to be finely tuned, too. A 'brute-force' approach that just tries to match an entire RNA in a single step will often cause CPU and memory usage to spin out of control. Therefore, stepwise strategies must be devised, which may be time consuming and complicated for non-experts. The ERPIN web server offers easy access to predefined motif searches. For each motif, the authors have developed and tested both the training sets and search parameters to achieve both fast database searches and satisfactory specificity/sensitivity ratios.

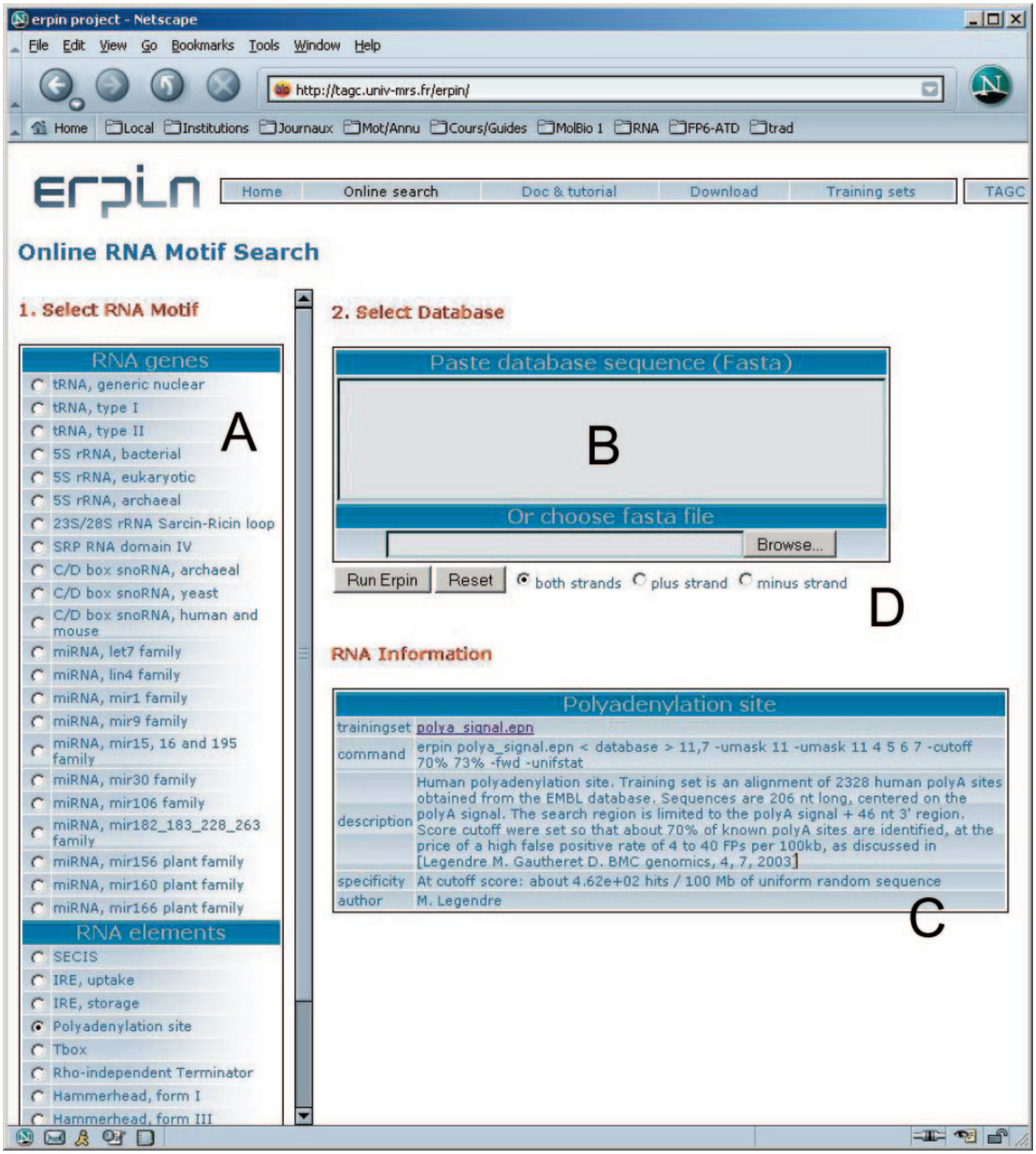
The current version of ERPIN running on the server is 4.03. Training sets and search parameters for each motif are stored in a MySQL database, accessed through a series of Perl and PHP scripts. Secondary structure graphics of solutions are produced by the NAVIEW program (8), adapted by Zuker and Ridgeway and obtained from M. Zuker's website (http://www.bioinfo.rpi.edu/~zukerm/export/). The web server is currently a dual-CPU 1.3 MHz Intel Xeon server with 2 Gb RAM.

The ERPIN search form is shown in Figure 1. Available RNA motifs or genes are shown in the left frame (Figure 1A). Selecting any gene or motif toggles the display of general information on this motif (Figure 1C), including training-set file, search parameters/command line, background information on the motif, training-set authors and available specificity information. Only one motif can be searched at a time. Once a motif is selected, users provide a database sequence to be scanned, either by copying it into the text field (Figure 1B) or by choosing a FASTA-formatted sequence file stored on their computer. Files containing multiple sequences are allowed as long as the total sequence size does not exceed 5 Mb. Finally, users may choose whether the search should be performed on the plus strand, minus strand or both (Figure 1D).

OUTPUT EXAMPLE

Predefined search parameters have been adjusted so that no run takes more than 3 min on a database the size of the *Escherichia coli* genome (both strands). Depending on server load, input sequence size and file transfer time, actual runtimes can vary between a few seconds and a few minutes. An example of a search result is shown in Figure 2, obtained after selecting type-II tRNA as the search motif and uploading the *Mycoplasma genitalium* genome (623 kb). A search on both strands runs in 1.95 s and produces nine solutions. For each solution, the server displays the name of the FASTA sequence in which the solution was identified (Figure 2A), the strand direction (Figure 2B), the position in the database sequence (Figure 2C), the score (Figure 2D), the *E*-value (Figure 2E) and, of course, the matching sequence itself. Gaps in matching sequences are inserted by the ERPIN algorithm so that each sequence is aligned with the initial training-set sequences. When certain regions of the training set are not used in the search, these are shown using lowercase letters.

The score (Figure 2D) is that computed directly from the lod-score profile. It does not convey much to a non-expert, unless it is compared with the cutoff score provided in the


erpin project - Netscape
File Edit View Go Bookmarks Tools Window Help
http://tagc.univ-mrs.fr/erpin/
Home Local Institutions Journaux Mot/Annu Cours/Guides MolBio 1 RNA FP6-ATD trad
erpin
Home Online search Doc & tutorial Download Training sets TAGC
A
B
C
D
Run Erpin
Reset
Browse...

Online RNA Motif Search

1. Select RNA Motif

RNA genes

○ tRNA, generic nuclear
○ tRNA, type I
○ tRNA, type II
○ 5S rRNA, bacterial
○ 5S rRNA, eukaryotic
○ 5S rRNA, archaeal
○ 23S/28S rRNA Sarcin-Ricin loop
○ SRP RNA domain IV
○ C/D box snoRNA, archaeal
○ C/D box snoRNA, yeast
○ C/D box snoRNA, human and mouse
○ miRNA, let7 family
○ miRNA, lin4 family
○ miRNA, mir1 family
○ miRNA, mir9 family
○ miRNA, mir15, 16 and 195 family
○ miRNA, mir30 family
○ miRNA, mir106 family
○ miRNA, mir182_183_228_263 family
○ miRNA, mir156 plant family
○ miRNA, mir160 plant family
○ miRNA, mir166 plant family

RNA elements

○ SECIS
○ IRE, uptake
○ IRE, storage
● Polyadenylation site
○ Tbox
○ Rho-independent Terminator
○ Hammerhead, form I
○ Hammerhead, form III

2. Select Database

Paste database sequence (Fasta)

Or choose fasta file

● both strands ○ plus strand ○ minus strand

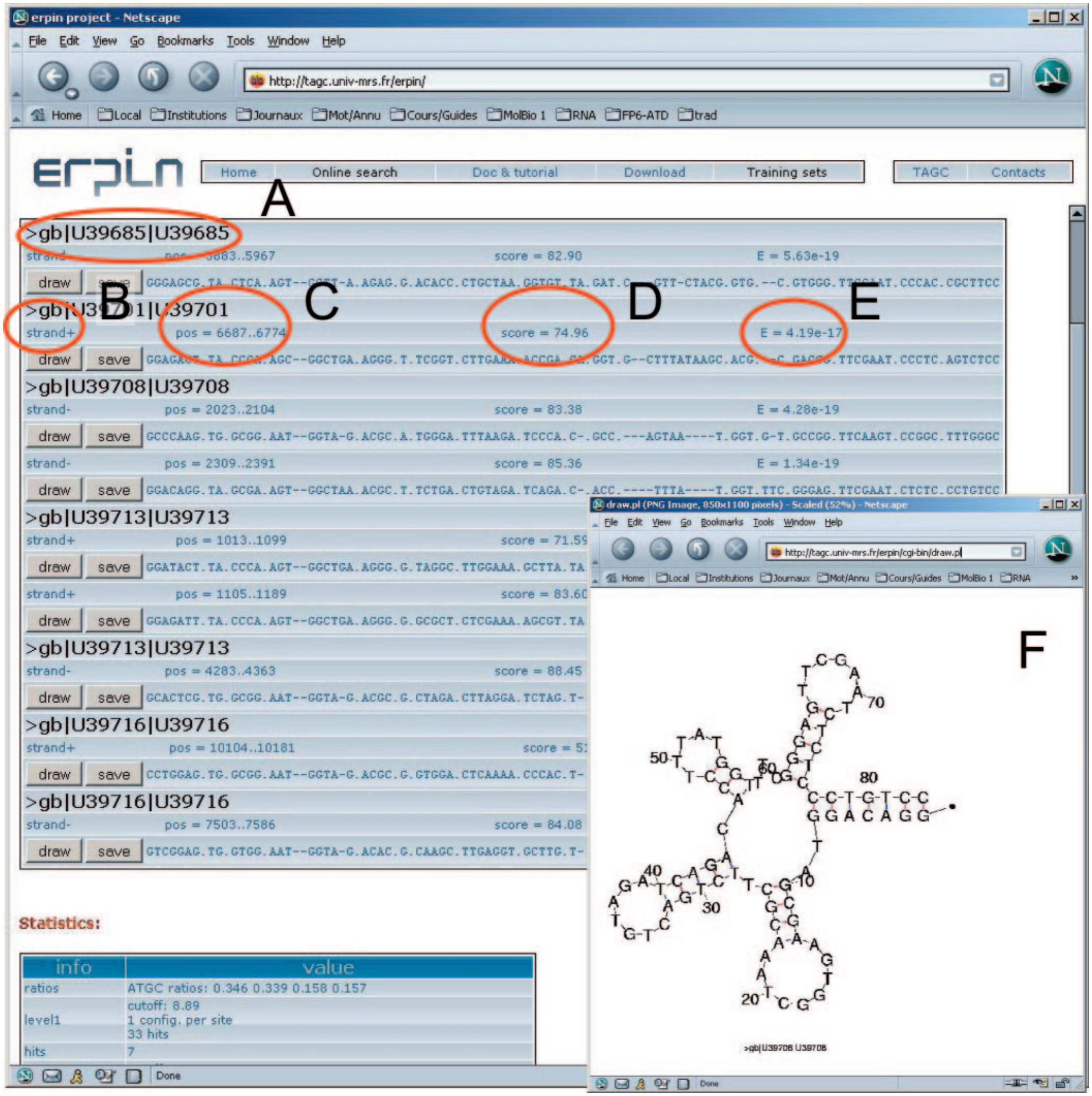
RNA Information

Polyadenylation site	
trainingset	polya_signal.epn
command	erpin polya_signal.epn < database > 11,7 -umask 11 -umask 11 4 5 6 7 -cutoff 70% 73% -fwd -unifstat
description	Human polyadenylation site. Training set is an alignment of 2328 human polyA sites obtained from the EMBL database. Sequences are 206 nt long, centered on the polyA signal. The search region is limited to the polyA signal + 46 nt 3' region. Score cutoff were set so that about 70% of known polyA sites are identified, at the price of a high false positive rate of 4 to 40 FPs per 100kb, as discussed in [Legendre M. Gautheret D. BMC genomics, 4, 7, 2003]
specificity	At cutoff score: about 4.62e+02 hits / 100 Mb of uniform random sequence
author	M. Legendre

Figure 1. Screen shot of the ERPIN server input form. The 'polyadenylation site' motif was selected and information related to this motif is displayed. (**A**) available motifs; (**B**) target database selection; (**C**) motif-related information; (**D**) selection of search strand.

'statistics' section at the bottom of the output page. When stepwise searches are performed, a distinct cutoff score is used at each step. Cutoffs cannot be changed by users of the web interface. For most motifs, cutoffs are set at the default value, which is the score of the lowest scoring training-set sequence. In some cases, however, cutoffs are modified in order to modulate sensitivity/specificity ratios. Such cases are discussed in the motif documentation. As explained above, *E*-values (Figure 2E) are much more informative than raw scores for the assessment of ERPIN hits. They depend both on score and database size. In our screen shot (Figure 2), *E*-values range from 10^{-17} to 10^{-19}, which is highly significant.

The 'draw' and 'save' buttons near each sequence respectively display and save to disk a secondary structure drawing of the RNA motif generated by the NAVIEW program (8). Base pairs in this drawing are based on the training set. Therefore, the structure is constant for a given RNA motif, and only the sequence varies. The type II tRNA training set used in this example has base pairs in the variable loop. This explains why the output


erpin project - Netscape
File Edit View Go Bookmarks Tools Window Help
http://tagc.univ-mrs.fr/erpin/
Home Local Institutions Journaux Mot/Annu Cours/Guides MolBio 1 RNA FP6-ATD trad
erpin
Home Online search Doc & tutorial Download Training sets TAGC Contacts
A
B
C
D
E
draw.pl (PNG Image, 850x1100 pixels) - Scaled (52%) - Netscape
File Edit View Go Bookmarks Tools Window Help
http://tagc.univ-mrs.fr/erpin/cgi-bin/draw.pl
Home Local Institutions Journaux Mot/Annu Cours/Guides MolBio 1 RNA
F
10
20
30
40
50
60
70
80
>gb|U39708 U39708
Done

```
>gb|U39685|U39685
strand-          pos = 5883..5967          score = 82.90          E = 5.63e-19
draw  save  GGGAGCG.TA.CTCA.AGT--GGTT-A.AGAG.G.ACACC.CTGCTAA.GGTGT.TA.GAT.C--GTT-CTACG.GTG.--C.GTGGG.TTCGAAT.CCCAC.CGCTTCC
>gb|U39701|U39701
strand+          pos = 6687..6774          score = 74.96          E = 4.19e-17
draw  save  GGAGAGT.TA.CCCA.AGC--GGCTGA.AGGG.T.TCGGT.CTTGAAA.ACCGA.GA.GGT.G--CTTTATAAGC.ACG.--C.GACGG.TTCGAAT.CCCTC.AGTCTCC
>gb|U39708|U39708
strand-          pos = 2023..2104          score = 83.38          E = 4.28e-19
draw  save  GCCCAAG.TG.GCGG.AAT--GGTA-G.ACGC.A.TGGGA.TTTAAGA.TCCCA.C-.GCC.---AGTAA----T.GGT.G-T.GCCGG.TTCAAGT.CCGGC.TTTGGGC
strand-          pos = 2309..2391          score = 85.36          E = 1.34e-19
draw  save  GGACAGG.TA.GCGA.AGT--GGCTAA.ACGC.T.TCTGA.CTGTAGA.TCAGA.C-.ACC.----TTTA----T.GGT.TTC.GGGAG.TTCGAAT.CTCTC.CCTGTCC
>gb|U39713|U39713
strand+          pos = 1013..1099          score = 71.59
draw  save  GGATACT.TA.CCCA.AGT--GGCTGA.AGGG.G.TAGGC.TTGGAAA.GCTTA.TA
strand+          pos = 1105..1189          score = 83.60
draw  save  GGAGATT.TA.CCCA.AGT--GGCTGA.AGGG.G.GCGCT.CTCGAAA.AGCGT.TA
>gb|U39713|U39713
strand-          pos = 4283..4363          score = 88.45
draw  save  GCACTCG.TG.GCGG.AAT--GGTA-G.ACGC.G.CTAGA.CTTAGGA.TCTAG.T-
>gb|U39716|U39716
strand+          pos = 10104..10181          score = 5
draw  save  CCTGGAG.TG.GCGG.AAT--GGTA-G.ACGC.G.GTGGA.CTCAAAA.CCCAC.T-
>gb|U39716|U39716
strand-          pos = 7503..7586          score = 84.08
draw  save  GTCGGAG.TG.GTGG.AAT--GGTA-G.ACAC.G.CAAGC.TTGAGGT.GCTTG.T-
```

Statistics:

info	value
ratios	ATGC ratios: 0.346 0.339 0.158 0.157
level1	cutoff: 8.89 1 config. per site 33 hits
hits	7

Figure 2. Screen shot of an ERPIN output page for a type II tRNA motif search. (**A**) sequence name in FASTA search database; (**B**) strand direction of hit; (**C**) position of hit; (**D**) score of hit; (**E**) *E*-value of hit; (**F**) a pop-up window showing the RNA structure, displayed by clicking on one of the 'draw' buttons.

structure (Figure 2F) has five stems instead of the usual four stems of tRNA. Although secondary structure displays are available for any motif, do not expect a consistent picture when no base-paired region is defined in the alignment (e.g. in polyadenylation sites) or when structures contain pseudoknots.

PERFORMANCE

At the time of submission, there were 22 genes and 12 shorter elements available for searching. For each of them, Table 1 provides a measure of search speed, specificity and sensitivity. Search speed is given as the CPU time required for scanning 1 Mb of uniform random sequence on both strands. In all cases, searches run in less than a minute per megabase (and most often in <10 s). Such performance allows us to offer large-scale searches of up 5 Mb even on a relatively small server. For larger searches (e.g. animal/plant chromosomes), users should submit their database in several parts or download the ERPIN program and training set for a local execution, using the same command as given on the website. Note that some training sets may not be available for public download, at the authors' request.

Table 1. Search performance of available RNA motifs

RNA motif	Time (s/Mb)[a]	FPs/100 Mb[b]	Sensitivity (%)[c]
tRNA, generic nuclear	7.88	9.81e − 04	97.84 (1086/1110)
tRNA, type I	3.71	1.04e − 01	98.34 (887/902)
tRNA, type II	1.38	6.09e − 07	94.69 (196/207)
5S rRNA, bacterial	4.98	1.21e − 10	92.59 (275/297)
5S rRNA, eukaryotic	4.98	1.46e − 29	88.92 (281/316)
5S rRNA, archaeal	4.23	8.61e − 14	96.30 (52/54)
23S rRNA, sarcin-ricin loop	3.76	1.74e − 01	86.15 (56/65)
SRP RNA domain IV	1.45	2.10e − 01	95.93 (165/172)
SECIS	22.91	6.38e + 01	82.79 (101/122)
IRE, uptake form	1.98	1.13e − 03	95.00 (19/20)
IRE, storage form	2.93	2.25e − 07	84.62 (11/13)
Polyadenylation site	0.44	4.62e + 02	63.52 (1478/2327)
T-box	26.58	1.14e + 01	92.14 (129/140)
snoRNA, C/D box archaeal	29.22	1.70e − 01	76.39 (165/216)
snoRNA, C/D box human–mouse	12.54	7.98e + 01	77.67 (80/103)
snoRNA, C/D box yeast	11.16	7.75e + 02	60.71 (17/28)
Rho-independent terminator	28.11	4.05e + 02	86.26 (1036/1201)
Hammerhead 1	28.13	1.90e + 01	100.00 (1657/1657)
Hammerhead 3	6.37	4.32e + 02	99.83 (2346/2350)
miRNA, let7 family	19.40	1.48e − 02	86.67 (26/30)
miRNA, lin4 family	5.81	9.84e − 02	81.82 (18/22)
miRNA, mir1 family	3.35	4.92e − 05	72.73 (8/11)
miRNA, mir9 family	26.23	1.22e − 01	72.73 (8/11)
miRNA, mir15-16-195 family	4.94	2.17e − 08	77.78 (7/9)
miRNA, mir30 family	3.58	5.24e − 03	93.33 (14/15)
miRNA, mir106 family	4.32	3.24e − 06	80.00 (8/10)
miRNA, mir182-183-228-263 family	4.05	1.53e − 02	75.00 (6/8)
miRNA, mir166 family, plants	31.47	6.26e − 02	85.71 (12/14)
miRNA, mir160 family, plants	6.72	8.99e − 07	62.50 (5/8)
miRNA, mir156 family, plants	12.64	7.66e − 09	71.43 (10/14)
CRE, poliovirus	4.06	9.03e − 28	100.00 (5/5)
IRES, poliovirus	3.09	1.02e − 194	100.00 (3/3)
TAR, HIV-1	2.33	2.60e − 09	100.00 (24/24)
PSTVd RNA, substructure	18.74	2.09e − 200	100.00 (89/89)

[a] Scan time for a 1 Mb random database, both strands, on a 1.3 GHz Intel Xeon server, excluding file transfer.

[b] Expected false positives per 100 Mb of random sequence (uniform nucleotide frequencies) at cutoff score, as computed by the ERPIN *E*-value function.

[c] Percentage of training-set sequences captured (number of captured training-set sequences/number of training-set sequences) using a 'leave one out' procedure (see text).

Specificity is usually expressed as the ratio TP/(TP + FP), where TP is the number of true positives and FP the number of false positives obtained during a search. This ratio varies with the search database, since it depends on the density of true sites in the database. Hence, it is common to use only a relative number of FPs (e.g. FPs per megabase) in a random sequence as a measure of specificity. Some motifs, however, are so specific that even a random sequence of 10 Gb or more would not be sufficient to find a single hit. Therefore, we use here the *E*-value computed by the ERPIN program to estimate the number of FPs that would be obtained at cutoff score in a 100 Mb random sequence of uniform nucleotide composition (Table 1).

Specificity ranges from excellent, with $<10^{-3}$ FPs per 100 Mb (5S rRNA, IRE or IRES) to relatively poor, with more than 10^2 FPs per 100 Mb (polyadenylation site, yeast C/D box snoRNA). Such variations are not surprising, as some RNA signatures are notoriously weaker than others. It is important, however, to emphasize that this only reflects FPs at *cutoff* score. This means that any solution with this score or better will be displayed, but most hits of biological significance are expected to come out with much higher scores, hence with much better *E*-values.

Sensitivity was computed using a 'leave one out' procedure, where each sequence is removed in turn from the training set and used as a search database with this reduced training set. The overall frequency and numbers of training set sequences captured are shown in Table 1. It should be of no surprise that sensitivity is satisfactory in most cases (>80%), given that this criterion was considered in the design of a proper command line for each motif. The lowest sensitivity is observed for motifs such as some miRNAs or C/D box snoRNAs, which are known to display a high signal-to-noise ratio. To reduce noise to an acceptable level, we had to tighten score cutoffs for these motifs, at the expense of sensitivity.

LIMITATIONS AND PERSPECTIVES

Although many RNA motifs are characterized by a well-defined signature, others are known to be elusive, either because they are poorly structured or constrained by base pairing to another RNA (e.g. box H/ACA snoRNAs) or because they vary too much. Even a well-structured motif such as the miRNA precursor can hardly be represented by a general model, since both the base-pairing pattern and the position of the conserved miRNA sequence vary considerably. In such cases, a practical search strategy generally involves identifying subfamilies (e.g. let-7 miRNAs) and constructing a model/alignment for each of them. Since the process of RNA sequence/structure alignment is still better carried out manually, the addition of a whole class of RNA such as miRNA or snRNA is a serious endeavor. Another limitation that is common to all RNA search programs is the difficult distinction between some pseudogenes or genomic repeats and evolutionarily related RNA genes, especially when divergence occurred recently. Finally, the current version of ERPIN does not handle insertions of arbitrary length at specific positions. This option, which is one of our development priorities, will enable the modeling of introns in genomic sequences. These caveats aside, the majority of known ncRNA genes or functional elements can be classified into subfamilies with specific and sensitive signatures that can be efficiently represented by lod-score profiles. Through a collaborative effort the ERPIN web server should soon cover the majority of these RNA motifs.

ACKNOWLEDGEMENTS

We thank the Action Bioinformatique Inter-EPST for its renewed support of the development of the ERPIN software and web server.

REFERENCES

1. Gautheret,D., Major,F. and Cedergren,R. (1990) Pattern searching/alignment with RNA primary and secondary structures: an effective descriptor for tRNA. *Comput. Appl. Biosci.*, **6**, 325–331.
2. Macke,T.J., Ecker,D.J., Gutell,R.R., Gautheret,D., Case,D.A. and Sampath,R. (2001) RNAMotif, an RNA secondary structure definition and search algorithm. *Nucleic Acids Res.*, **29**, 4724–4735.
3. Billoud,B., Kontic,M. and Viari,A. (1996) Palingol: a declarative programming language to describe nucleic acids' secondary structures and to scan sequence database. *Nucleic Acids Res.*, **24**, 1395–1403.
4. Eddy,S.R. and Durbin,R. (1994) RNA sequence analysis using covariance models. *Nucleic Acids Res.*, **22**, 2079–2088.
5. Griffiths-Jones,S., Bateman,A., Marshall,M., Khanna,A. and Eddy,S.R. (2003) Rfam: an RNA family database. *Nucleic Acids Res.*, **31**, 439–441.
6. Gautheret,D. and Lambert,A. (2001) Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles. *J. Mol. Biol.*, **313**, 1003–1011.
7. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
8. Bruccoleri,R.E. and Heinrich,G. (1988) An improved algorithm for nucleic acid secondary structure display. *Comput. Appl. Biosci.*, **4**, 167–173.